DOCUMENT RESUME

ED 194 560 TM 800 689

TITLE The National Item Analysis Program. Metropolitan Achievement Tests, 1978 Edition. Special Report Number 7.
INSTITUTION Psychological Corp., New York, N.Y.
PUB DATE 78
NOTE 17p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests: Elementary Secondary Education: *Item Analysis: *Participant Characteristics: *Sampling: Standardized Tests: *Test Construction: *Test Items
IDENTIFIERS *Metropolitan Achievement Tests

ABSTRACT

The following information on the development of the 1978 Metropolitan Achievement Tests (MAT), fifth edition, is given: the purpose and importance of item analysis, steps preceding item tryout, characteristics of the sample population, and data obtained. Item analysis allows elimination of undesirable items, adjustment of the difficulty level of the test, and production of tests with known approximate reliabilities. Effective item analysis depends on the number of items tried out and the characteristics of the student sample. Activities preceding item analysis include content development, item writing and editing. The sample population for the current MAT included approximately 93,000 students from kindergarten through grade 12. Selected schools represented the national population by socioeconomic level and geographic distribution. For each item in the item analysis, the following data were obtained: the difficulty and discrimination of the item summarized by grade and a breakdown of data for the item at each grade level tested. Data for complete tests included frequency distributions of raw scores, Kuder-Richardson Formula 21 reliabilities, distributions of item difficulties and item discrimination indices, and intercorrelations among tests. These data were also obtained separately for each grade taking each form of a given battery and level. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED194560

Metropolitan Achievement Tests

1978 Edition

Special Report NUMBER 7

The National Item Analysis Program

The development of the Metropolitan Achievement Tests involved two major research programs--the national item analysis program and the standardization program. Details of the standardization program are presented in Special Reports 8, Guidelines for Standardization Sampling, and 9, Summary Statistics for National Standardization Groups. This Report describes the purpose and importance of item analysis, the steps preceding the item tryout, and the methods and results of the study: the characteristics of the sample selected, and the data obtained. Special Reports 1, Content Development, and 4, Content Validity, provide a detailed account of the steps preceding the item analysis program.

Contents

I The New Series

II Purpose and Importance of Item Analysis

III Steps Preceding Item Analysis

A. Content Development

B. Item Writing and Editing

IV General Information

V Characteristics of the Sample

A. Sample Selection

B. Sample Verification

VI Data Obtained from the Program

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Psyc. Corp.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Psychological Corporation
A subsidiary of Harcourt Brace Jovanovich, Inc.

Copyright © 1978 by Harcourt Brace Jovanovich, Inc. All rights reserved. Printed in U.S.A.

TM 800689

I The New Series

In 1973, work on the fifth edition of the Metropolitan Achievement Tests was begun. It has become the practice at The Psychological Corporation to revise the major achievement tests every eight to ten years in order to ensure that the tests reflect curriculums established in the nation's schools and that the normative data remain current, and to utilize significant advances in the field of test construction.

The 1978 Metropolitan incorporates a two-component system of achievement testing. The Survey component provides efficient assessment in all major curricular areas; the Instructional component was designed for in-depth criterion-referenced assessment in the basic areas of reading, mathematics, and language arts. The Instructional batteries are based on instructional objectives, each assessed by at least three test items. The purpose of the national item analysis program was to collect data on students' performance on the individual items and on the way the several items testing the same objective functioned together.

II Purpose and Importance of Item Analysis

An item analysis program is conducted to obtain data on proposed items for a test. Analysis of these preliminary data allows selection of optimum sets of items for the final test. The focus of attention in the item analysis program is the individual items. Complete tests and batteries of tests become the major focus of attention in the standardization program. The item analysis is obviously an exceedingly critical prelude to standardization and ultimate use of the tests, because analysis of individual item data allows the test developer to eliminate items with undesirable characteristics, to adjust the final difficulty level of the tests as needed, and to produce for

standardization tests with known approximate reliabilities. In many respects, the item analysis program for a test is more important than the standardization. The standardization program determines the test norms; these may not be used at all by schools using local norms or using only criterion-referenced scores on individual objectives. On the other hand, the item analysis program to a large degree determines the test content, which affects everyone using the test.

The effectiveness of the item analysis program depends on two factors: the number of items tried out and the characteristics of the sample of students tested. Use of more items in the program than will be needed for the final forms of the test adds greatly to the value of the item analysis. This allows the test developer to try out novel items, to try out items that approach the same objective in slightly different ways, to try out items on the same objective at varying levels of difficulty, and to try out different items assessing the same objective at several battery levels. These variations provide the test developer with flexibility in combining optimum sets of items for the final tests.

The characteristics of the item analysis sample are also an important consideration. The sample should be large enough to yield fairly stable data on the many items tried out. It should include representative schools from all geographic regions with a variety of socioeconomic and school characteristics. The sample should, on the average, represent the national school population in characteristics which are related to test performance.

III Steps Preceding Item Analysis

A. Content Development. Content development was the first stage in the creation of the new Metropolitan. The authors and editorial staff surveyed

state curricular guides and syllabuses, and, most importantly, the most widely-used textbooks in all content areas. Analysis of these sources resulted in such decisions as what objectives to include at each battery level, typical modes for presenting various topics, what balance of topics to include in the various content areas (e.g., in science, more biological science than physical science items; in social studies, fewer geography than history items), etc.

B. Item Writing and Editing. After test blueprints reflecting these and other decisions were developed, reviewed, and finalized, actual item writing began. The vast majority of items (over 95%) on the Metropolitan were newly-written for this edition. Item writing was done by the test authors, each of whom was responsible for one or more domains (content areas). The authors performed various small-scale tryouts of items in local school systems, in order to verify judgments of pupil performance on various types of items, and of the difficulties of specific items and item formats at various grade levels.

Complete sets of items were written and forwarded to the editorial staff of The Psychological Corporation. All items tried out on the item analysis edition were put through a five-stage editorial process. First the items were edited by specialists in the curricular area being tested. The primary focus of attention was on correctness and appropriateness of item content. Second, the items were edited by measurement experts whose major concern was the application of good item-writing techniques. Third, the items were edited by a group of general editors. Their responsibility was the wording of items, the grammatical structure, and related matters. Fourth, the items were reviewed and edited by a panel of minority educators, whose primary concern was the appropriateness of the content for children of both sexes and of

various cultural groups.* Finally, all teacher involved in the item analysis program were asked to comment on the items, on the tests as a whole, and on the directions, thus constituting a fifth group who shared in the editing process.

IV General Information

The Metropolitan Item Analysis (IA) program took place in two phases: Phase I, October 13 to 31, 1975; and Phase II, January 12 to 30, 1976. Approximately 93,000 pupils in Grades K through 12 from 44 school systems in 35 different states were tested. Each pupil took the appropriate level of the Word Knowledge Test (Form H) of the 1970 Metropolitan, and one form of one of the two experimental batteries of the new Metro, either Reading/Language or Mathematics/Science/Social Studies. Since the item analysis tests were too long to ask each system to test in all content areas, two separate samples of school systems were drawn and assigned to either the Reading/Language or Mathematics/Science/Social Studies groups.

In Phase I, 20 participating systems were tested with the Reading/Language Tests and 22 systems took the Mathematics/Science/Social Studies Tests. All systems tested in Grades 1-6 or 1-7. In Phase II, 10 systems tested in Reading and 11 in Mathematics. Most Phase II systems were asked to test in Kindergarten and in some combination of grades 7-12. Additionally, 5 Phase II systems tested in Grade 6.

All pupils in a system were included in the testing, except in 16 very large systems where samples known to be representative of the entire system in terms of previous test scores and ethnic population were tested.

---

*See Special Report 16 for further information about the role of the Advisory Panel.

Approximately 13,000 new test items, divided among three experimental forms at each of seven battery levels, were tried out. The three forms of each level were distributed randomly by class throughout participating systems. In general, each battery was tried out one grade above and one grade below as well as at the grade on grades for which it was finally intended. Table 1 shows grades tested with each battery in the IA program.

Table 1

Grades Tested in the Item Analysis Program

by Battery, Testing Phase, and Form

BATTERY	TESTING PHASE	GRADES	FORMS
Primer	I,II	K.4, 1.1, 2.1	1,2,3
Primary 1	I	1.1, 2.1, 3.1	1,2,3
Primary 2	I	2.1, 3.1, 4.1	1,2,3
Elementary	I	3.1, 4.1, 5.1	1,2,3
Intermediate	I	4.1, 5.1, 6.1, 7.1	1,2,3
Advanced 1	II	6.4, 7.4, 9.4, 11.4	1,2,3
Advanced 2	II	8.4, 10.4, 12.4	1,2,3

The Word Knowledge (Vocabulary) Test from the 1970 edition of the Metropolitan Achievement Tests was used as a control instrument to check the representativeness and balance of the general achievement level of the sample across forms, grades, and test levels.

V Characteristics of the Sample

The effectiveness of an item analysis program depends on the nature of its sample. A good sample is representative of the students who eventually will use the final test. To this end, information was collected on a variety of characteristics pertaining to the sample of students. Some of this information was used to help select the sample; the remainder was used to verify the sample's appropriateness.

The first part of this section describes the use of sample characteristics in selection of the sample. The second part compares the sample and the national data used for verification purposes.

A. Sample Selection. Schools in the IA sample were selected to provide appropriate representation of the national population according to socio-economic level and distribution by geographic region. Of these two variables socio-economic level was considered the more critical, since it relates highly to test performance.

Application of demographic data such as those available from the U.S. Census Bureau reports always involves some inaccuracy, because boundaries of school systems and boundaries of political units such as cities often differ. Exactly which demographic figures are accurate in these situation is problematic.

To help correct for this problem, a weighting procedure was used to derive more accurate figures describing the populations of the sample school districts. Census figures for all demographic areas cutting into a single school district were weighted and combined. The weights were based on data obtained from each IA school system. For example, if 25% of the children in the Oak Park District came from Park County, Park County figures would comprise 25% of the final composite figures for the Oak Park School District.

Two socioeconomic variables were used to define the IA sample: median family income (MFI) and median years of school (MYS) completed by persons aged 25 years and over in the sample communities. Data for these two characteristics in cities and counties are available in the 1970 U.S. Census Bureau records. Table 2 presents summary statistics for the IA sample and for the U.S. population on income and years of schooling. The sample means and medians were obtained by weighting figures for each system or community by the number of

pupils tested.

Comparisons of sample medians for years of schooling with the national figure show a very close match between the figures, although slight differences in the mean values indicate that the sample groups differ from each other somewhat. Comparison of the income figures indicates that the sample communities tend to be slightly poorer than the nation as a whole. Furthermore the sample groups seem to differ from each other somewhat in terms of family income.

TABLE 2

Summary of Socioeconomic Characteristics* of School Systems Participating in the National Item Analysis Program - October, 1975 & January, 1976

PROGRAM	Content Sample	Median Years Schooling		Median Family Income (in hundreds)	
		Median	Mean	Median	Mean
PHASE I October, 1975	Reading/Language	12.1	11.3	$ 95	$ 92
	Mathematics/ Science/Scoial St.	12.1	11.6	92	98
	TOTAL	12.1	11.5	93	95
PHASE II January, 1976	Reading/Language	12.1	11.1	95	80
	Mathematics/ Science/Social St.	12.1	11.8	88	96
	TOTAL	12.1	11.5	93	89
NATIONAL FIGURES **		12.1	-	$ 96	-

*Data are weighted by approximate sample sizes within each system.
**Based on 1970 U.S. Census Bureau figures.

Table 3 shows representation of the IA sample by geographic region. The regional breakdown of states employed here is that adopted by the National Education Association.

TABLE 3

Regional Breakdown of the Item

Analysis Sample and National Population

PHASE I

Region	% Total US Population	Rdg/Lang		Math/Sci/SS		Total Phase I Sample	
		No.of Systems	%	No.of Systems	%	No.of Systems	%
Northeast	26	5	32	3	17	8	24
North Central	29	5	31	9	36	14	34
South	20	5	19	3	18	8	18
West	24	5	18	7	28	12	24

PHASE II

Region	% Total US Population	Rdg/Lang		Math/Sc/SS		Total PhaseII Sample	
		No.of Systems	%	No.of Systems	%	No.of Systems	%
Northeast	26	2	39	3	34	5	36
North Central	29	1	4	5	39	6	25
South	20	4	30	2	9	4	17
West	24	5	27	3	19	6	22

Since all of the pupils in the IA program took the MAT'70 Word Knowledge Test it was possible to describe the sample in terms of national percentile ranks on this test. Tables 4a and 4b give the mean percentile ranks across forms for all of the grade levels tested. The norms for the students who took the Advanced 2 and Primer batteries are estimated, since Metro'70 did not include these levels. All percentile ranks are in the 41-60 range. The means and medians approximate 50 across grades. In Grades 2-9, where empirical norms were available for comparisons, all means were between 47 and 55.

Mean Word Knowledge National Percentile Ranks

For Item Analysis Program Sample

TABLE 4a — READING/LANGUAGE SAMPLE

Test Level and Form

	Primer			Primary 1			Primary 2			Elementary			Intermediate			Advanced 1			Advanced 2		
Grade	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
K.4*	54	54	53																		
1.1*	55	55	54	54	53	54															
2.1	54	53	53	52	52	52	53	52	52												
3.1				52	53	53	52	51	51	51	51	50									
4.1							51	51	51	48	48	49	48	48	48						
5.1										49	49	50	49	49	48						
6.1/.4													48	48	49						
7.1/.4													48	48	48	52	53	54			
8.4																			51	50	50
9.4																48	48	47			
10.4*																			52	52	51
11.4*																46	41	44			
12.4*																			59	59	60

TABLE 4b — MATH/SCIENCE/SOCIAL STUDIES SAMPLE

Test Level and Form

	Primer			Primary 1			Primary 2			Elementary			Intermediate			Advanced 1			Advanced 2		
Grade	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
K.4*	57	56	56																		
1.1*	55	55	55	54	56	54															
2.1	52	52	52	52	53	52	52	53	52												
3.1				50	50	50	52	52	52	53	53	53									
4.1							50	49	50	51	51	51	50	50	50						
5.1										50	50	50	50	50	49						
6.1/.4													51	51	51	55	55	55			
7.1/4													49	49	49	52	52	52			
8.4																			**		
9.4																52	52	52			
10.4*																			**		
11.4*																59	59	59			
12.4*																			**		

*Estimated percentile ranks

**Advanced 2 data are listed in the Reading/Language table.

The characteristics of the IA sample all appear to be satisfactory. The sample distributions are generally comparable to national figures. In light of the data obtained on the MFI and MYS indices and on performance on the Word Knowledge Test, it was decided that post-hoc weighting of the sample to improve representation was unnecessary.

B. Sample Verification The data described in the preceding section were used in the sample selection process. Data were also collected for verification purposes. Comparison of characteristics of the sample with characteristics of the national school population help ensure that the sample is truly representative. It is important to know that both the students in the sample and the schools attended by those students are representative. Therefore two types of data were collected and used: characteristics pertaining to the students and characteristics of the school systems from which the sample was drawn.

Most of the data describing the sample were derived from responses on a questionnaire filled in by a representative of each school system participating in the program. Student characteristics provided by this source include ethnic composition, urban-rural composition, and percent of first graders that had attended kindergarten. School system characteristics include starting teacher salary, average teacher salary, average per pupil expenditure, pupil-teacher ratio, and average years of teacher experience. The results derived from these questionnaires are presented in this section, along with the national population figures used for comparison. In addition, school enrollment figures are included. A final section presents some figures which describe the sample, but for which no national comparisons were made.

Table 5 shows data for all IA schools combined and for 1970 Census summary data on ethnic composition of the student population. The designations used in the table (i.e. White, Black, Hispanic, Oriental, American Indian) were selected because they are commonly used by both government and social research surveys. They are also used and understood by the education community at large. In general, the IA sample matches U.S. data quite closely.

TABLE 5

Ethnic Breakdown

for IA Sample and U.S. Population

Ethnic Designation	% Total IA Sample	% Total Population (1970 Census)
White	82.7	83.9
Black	11.9	11.1
Hispanic	3.5	4.6
Oriental	.7	.01
American Indian	1.2	.4

The urban-rural distribution of the IA sample closely matches that of the U.S. population as well. On the average, IA systems reported that approximately 24% of their pupils come from basically rural areas. The national population is 26% rural (non-urban).

TABLE 6

Rural/Urban Population Distribution for

IA Sample and U.S. Population

	% Total IA Sample	% Total U.S. Population
Urban**	75.9	73.5*
Rural	24.1	26.5

*1970 Census Data
**Includes areas surrounding large cities, towns, and villages.

For the student characteristics reported above, public and private schools were combined. For some characteristics, however, these two types of schools should be distinguished, as national figures are available only for public schools. Table 7 presents the comparison between public IA school systems and the national school population for some of these characteristics: starting teacher salary, per pupil expenditure, pupils per teacher (pupil-teacher ratio) and average years of teaching experience for teachers. Column 2 of Table 7 presents mean data for public schools in the IA sample. Column 4 gives mean data for the total IA sample, i.e., for public and parochial schools combined. The data in Column 2 should be used in comparing the IA sample with the U.S. figures in Column 1. Columns 3 and 5 provide median data. In general, it appears that the IA sample closely approximates the national population on these variables. The ranges for each characteristic in Table 7 for public schools in the IA sample are as follows: starting teachers' salary

$5500 to $10,500; average teachers' salary, $6500 to $14,500; per-pupil expenditure, $300 to $1800; pupils per teacher, 19 to 33; average years' teaching experience for classroom teachers, 5 to 14. Considering the variability of these characteristics, differences between the IA sample and the national population appear to be negligible.

TABLE 7

Comparison of Selected School Characteristics for the IA Sample (Public Schools only and Total Sample) and U.S. School Population

Characteristic	U.S. School Population	IA Public Schools		IA Total Sample	
		Mean	Median	Mean	Median
Starting Teacher Salary	$7500[a]	$ 8021	$ 8332	$ 7740	$ 8000
Average Teacher Salary	$11513[b]	$11165	$10776	$10563	$10437
Per Pupil Expenditure	$1255[c]	$ 993	$ 904	$ 923	$ 852
Pupils Per Teacher	20.7[d]	26.0	25.7	26.1	25.5
Average Yrs. Teaching Experience	8[e]	10	11	11	11

a NEA Research Bulletin, 1973-74.
b NCES Digest of Educational Statistics, 1975 ed., Table 50 (data for 1974-1975).
c Ibid., Table 71 (data for 1974-1975, day schools only).
d Ibid., p. 2.
e Ibid., Table 47 (data for 1970-1971).

School system enrollment was not used as a stratifying variable in selecting the IA sample. However, questions are sometimes raised about representation on this variable. Comparisons of the U.S. school population and the Metropolitan IA sample, all school systems and public schools only, in terms of school enrollment are shown in Table 8. Although there is some unevenness in the IA distribution compared to the national distribution, representation on this variable seems to be satisfactory.

TABLE 8

Comparison of the National School Population and Metropolitan Item Analysis Sample in Terms of School System Enrollment

Total School Enrollment	% Total IA Sample	% IA Public School Sample	% U.S. School Population*
25,000 +	30	32	29
12,000 - 24,999	7	3	15
6,000 - 11,999	14	14	17
3,000 - 5,999	20	19	17
1,800 - 2,999	24	30	10
600 - 1,799	5	3	10
0 - 599	0	0	3

*Table 62, NCES Digest of Education Statistics, 1974 ed., (October, 1971 enrollment figures).

The survey of the sample systems generated some additional information that may be of interest to the test user. About 5% of the students in the IA sample spoke a language other than English outside of school, or came from homes in which a language other than English is spoken most of the time. Twenty-five of the sample systems (66%) reported using homogeneous class groupings, usually in all grades, and based on some combination of teacher

recommendations, standardized test results, and grades.

VI Data Obtained from the Program

An example of the type of data obtained for each item in the item analysis program is shown on a sample item card in Figure 1. One such card was prepared for each item, and all data relating to the item, plus the item itself and teacher's directions if necessary, were affixed to the card. The upper left-hand part of the card contains identifying information and the item itself. In the lower left portion of the card, the difficulty and discrimination of the item is summarized by grade. The right-hand side of the card shows the breakdown of data for the item at each grade tested.

All of the above data pertain to individual items. In addition, certain data for complete tests were obtained. These included frequency distributions of raw scores, Kuder-Richardson Formula 21 reliabilities, distributions of item difficulties and item discrimination indices, and intercorrelations among tests. These data also were obtained separately for each grade taking each form of a given battery and level.

FIGURE 1